Abstract Title Page
Not included in page count.

Title: Extensions of existing methods for use with a new class of experimental designs useful when there is treatment effect contamination

Author(s): Christopher Rhoads, PhD, Northwestern University

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Researchers planning a randomized field trial to evaluate the effectiveness of an educational intervention often face the following dilemma. They plan to recruit schools to participate in their study. The question is, "Should the researchers randomly assign individuals (either students or teachers, depending on the intervention) within schools to treatment conditions, or should all participating students at a given school be assigned to the same treatment condition?". That is, should we randomize schools (clusters), or individuals within schools?

One reason often given for preferring cluster level randomization is a fear of "diffusion of treatment" (Raudenbush, 1997) or "contamination" (Donner and Klar, 2000). Contamination occurs when contact between members of the control group and members of the experimental group causes control group participants to behave more like experimental group participants than they would have had that contact not occurred. It is also possible for certain interventions that contamination could cause experimental subjects to behave more like control group subjects than they would otherwise. Note that both experimental subjects acting more like control subjects and control subjects acting like experimental subjects are processes that would tend to decrease the effect size of the experiment. I assume for the purposes of this paper that the contamination dilutes the observed effect size in this fashion. The methods considered would not apply to a contamination process that tends to spuriously increase the effect size, such as control group demoralization (Shadish, Cook & Campbell, 2002).

Cornfield (1978) noted that two penalties are paid for randomization by cluster rather than by individual. First, the variance of the estimated treatment effect increases. Second, the degrees of freedom available to estimate that variance decrease. Thus, in the absence of contamination, randomizing an equal number of individuals within each cluster to each treatment (often called a *randomized block* or RB design) is a more powerful design than randomizing whole clusters (often called a *cluster randomized* or CR design). Rhoads (forthcoming) has argued that the threat of contamination should not necessarily lead experimenters to opt for a cluster randomized design. He points out that, depending on the values of relevant design parameters (i.e., the ICC, within cluster sample size, the heterogeneity in treatment effects across clusters, and the number of clusters in the experiment) the statistical power of a randomized block design remains higher than the power of a cluster randomized design even when contamination causes the effect size to decrease by as much as 10-60%. Similarly, from the standpoint of mean squared error, Rhoads (forthcoming) shows that for many design parameters of practical interest the randomized block design will be preferred to the cluster randomized design.

However, it may well be the case that the optimal experimental design in the presence of contamination is neither the RB design nor the CR design, but some compromise between the two designs. This is precisely the situation considered by Borm, Melis, Teerenstra and Peer (2005) (hereafter BMTP) who suggest an interesting compromise between cluster randomization and individual randomization, a method they call "pseudo cluster randomization." They label the two treatments that are under investigation as treatments A and B. Suppose that 2m clusters,

each of size *n,* are available for the experiment. Pseudo cluster randomization is a two step randomization procedure. First, the clusters are randomly assigned to two groups labeled *a* and *b*, resulting in *m* clusters per group. Then, for each cluster in group *a*, a fraction $f\ (0.5 \le f \le 1)$ of the individuals in that cluster are randomly assigned to treatment A and the rest to treatment B. In cluster group *b*, the same fraction *f* of individuals in each cluster are assigned to treatment B and the rest to treatment A. Using the same fraction *f* in each cluster group ensures that the design is balanced on both the individual and the cluster level.

Next, define $y_{\mathrm{A}a_j}$ and $y_{\mathrm{A}b_j}$ to be the mean outcome of individuals assigned to treatment A in cluster groups *a* and *b*, respectively. Notation for the mean outcomes of those receiving treatment B are defined analogously. It may be advantageous to weight responses in the different cluster groups differently. Thus, BMTP propose the following estimator of the mean outcome for those receiving A

$$y_{\mathrm{A}} = \frac{1/m \sum f\, y_{\mathrm{A}a_j} + 1/m \sum w(1-f)\, y_{\mathrm{A}b_j}}{f + w(1-f)}. \tag{1}$$

Then the variance of the estimated treatment difference $y_{\mathrm{A}} - y_{\mathrm{B}}$ is

$$\mathrm{var} = \frac{2\sigma^2}{mn} \frac{f + w^2(1-f) + nq\{f - w(1-f)\}^2}{\{f + w(1-f)\}^2}, \tag{2}$$

where *q* is defined by $q = \rho/(1-\rho)$, and ρ is the intracluster correlation coefficient.

Contamination is conceptualized as follows. Let μ_A and μ_B be population mean outcomes under treatments A and B in the absence of contamination and let $\delta = \mu_A - \mu_B$. Then the expected outcome for those receiving treatment A in group *a* is $\mu_A - c_{f,A}\delta$ $(0 \le c_f \le 1)$. This is the contamination of A by B given that a fraction, *f,* receive A in each cluster. Defining other contamination factors in an analogous fashion we find that the expected value of $y_A - y_B$ with contamination is given by

$$E = \delta - \delta \frac{f(c_{f,A} + c_{f,B}) + w(1-f)(c_{1-f,A} + c_{1-f,B})}{f + w(1-f)}. \tag{3}$$

BMTP note that the ratio $t^2 = E^2/\mathrm{var}$ is inversely proportional to the number of subjects required to achieve fixed type I and type II error rates using *z* critical values. Hence, from the standpoint of statistical power one would prefer experimental designs with larger values of t^2.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

This research focuses on extending results found in BMTP (2005) to the following cases: (1) the case where the value of the ICC cannot be estimated precisely, (2) the case where treatment

effects vary across clusters, (3) the case where the criterion function for determining an optimal design/estimator is mean squared error, rather than statistical power.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

There are three major limitations to the work done by BMTP that the current work overcomes. First, BMTP assume a large number of clusters enrolled in the experiment and that the value of ρ is known. The current work explores this issue. That is, it looks at what can be done when the value of ρ cannot be estimated precisely. Second, the BMTP work assumes homogeneous treatment effects across clusters. This assumption appears to be standard in the health sciences literature, however, it is not recommended in educational experiments where it is possible, if not likely, that the effect of treatment may vary across different educational settings. Third, BMTP assume that the researcher will always prefer the design with more statistical power. However, much recent work has recognized that researchers make a serious mistake when they summarize the results of their work solely in terms of the results of significance tests (Schmidt and Hunter, 1997; Ziliak and McCloskey, 2004). Hence, this paper compares the RB and CR designs both using statistical power as the criterion for deciding between the designs and using the mean squared error of the point estimate of the average treatment effect as the criterion. The results section will show that these two criteria do not always lead to the same decision.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

The statistical model assumed throughout the paper is as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, i = 1,2; j = 1,\ldots,2m; k = 1,\ldots, n_{ij}, n_{1j} + n_{2j} = n. \quad (4)$$

where μ and α_i are fixed parameters representing, respectively, the overall mean and the deviation of the mean in treatment group *i* from the overall mean, β_j is mean zero normally distributed random effect associated with cluster *j* having variance σ_B^2, $(\alpha\beta)_{ij}$ is a mean zero normally distributed random effect representing the interaction of the treatment effect with the cluster effect and having variance σ_{TC}^2, and ε_{ijk} is a mean zero, normally distributed random effect associated with individual *k* within cluster *j* and has variance σ^2. Many results will depend on two relevant summary quantitites. First, the *intracluster correlation coefficient* (ICC), defined as $\rho = \sigma_B^2/(\sigma_B^2 + \sigma^2)$. Second, the parameter $\omega = \sigma_{TC}^2/\sigma_B^2$, which represents the ratio of one half of the variation in treatment effects across clusters to the total variation across clusters.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

I believe that the usefulness of the methods is evident from the results presented in the next section.

Findings / Results:
Description of the main findings with specific details.

There are three distinct types of results, corresponding to the three ways in which this work generalizes the work of BMTP. First, I present results comparing the RB design and the CR design when mean squared error (MSE) is used as the criterion to choose between these designs. Results are contrasted with the results that would be achieved if statistical power were used as the criterion. Results are reported in terms of MAC, or *maximum allowable contamination*, which is defined as the amount of contamination that could be tolerated before a CR design would be preferred to a RB design. When statistical power is used as the criterion, the formula defining maximum allowable contamination is

$$\mathrm{MAC}_{p,het}(n,\rho,\omega) = 1 - \sqrt{\frac{1+(n\omega/2-1)\rho}{1+(n-1)\rho}}\,. \tag{5}$$

When MSE is used as the criterion, the MAC formula is given by

$$\mathrm{MAC}_{MSE,het}(m,\rho,\omega,\delta_T) = \sqrt{\frac{2\rho}{m}}\frac{1}{\delta_T}\sqrt{1-\frac{\omega}{2}}\,. \tag{6}$$

The quantity δ_T is defined as $\delta_T = \delta / \sqrt{\sigma_{TC}^2 + \sigma_B^2 + \sigma^2}$. Tables 1 and 2 evaluate the formulae given in equations (5) and (6) for various values of the relevant design parameters. Table 1 presents results for the case ω=0 (homogeneous treatment effects), and table 2 presents results for various positive values of ω. The first three rows of each table present the corresponding results when statistical power is used as a criterion. The tables demonstrate that these two criteria can lead to very different results.

The next set of results look at what sort of results can be obtained when pseudo cluster randomization is used but the value of the ICC cannot be assumed to be known. I note from equation (2) that if *w=f*/(1-*f*) is used as the weighting function to form the estimator of the average treatment effect, then the variance of the estimated treatment effect does not depend on ρ. This is useful because frequently σ^2 can be estimated with much more precision than ρ. I will refer to a treatment effect estimate constructed with weights *w=f*/(1-*f*) in as a *weighted-invariant* (WI) estimator. The expected value of the estimated treatment effect for the WI estimator will depend on the average contamination in each treatment at fractional randomization *f*, $c_{fa} = (c_{f,A} + c_{1-f,A} + c_{f,B} + c_{1-f,B})/2$. While BMTP allow all values of *c* to vary from 0 to 1 in an unrestricted fashion, this would imply that contamination could cause all control subjects to behave as though they were *uncontaminated* experimental subjects and all experimental subjects to behave as though they were uncontaminated control subjects, resulting in the absolute value of δ staying the same but δ changing sign. This seems implausible, and so the restrictions $c_{f,A} + c_{1-f,B} \le 1$ and $c_{f,B} + c_{1-f,A} \le 1$ seem logical. The degrees of freedom of the WI estimator are the same as the degrees of freedom for an RB design, so a very good approximation to the relative sample size required under the two designs is given by

$$\frac{t_{wi}^{-2}}{t_{1:1}^{-2}} = \frac{(1-c_{0.5a})^2}{4f(1-f)(1-c_{fa})^2}\,. \tag{7}$$

Since $f(1-f)$ is at most 0.25, in the absence of contamination we obtain the well known result that a balanced experimental design is the most efficient. The WI design can improve on the 1:1 design only if we can remove a substantial amount of contamination by using $f > 0.5$. For instance, suppose that we believe that experimental subjects are unlikely to be contaminated by control subjects and that at 1:1 randomization the control group mean will move halfway towards the experimental group mean. That is, we think $c_{0.5,A} = 0$ and $c_{0.5,B} = 0.5$. Now suppose that at a fractional allocation f=0.8 experimental subjects are still uncontaminated by control subjects, when 80% of the subjects in a cluster are in the experimental group contamination of controls increases to $c_{0.2,B} = 0.6$, however when only 20% are in the experimental group, contamination decreases to $c_{0.8,B} = 0.1$. Then equation (6) results in 0.925 and the WI design would be preferred. On the other hand, if $c_{0.8,B} = 0.2$ the 1:1 design is preferred.

The WI estimator will be preferred only if the right hand side of equation (7) is less than one. Hence, given an anticipated amount of contamination in the RB design, we can determine how much less average contamination would need to be in order to want to use a fraction allocated to treatment that is other than ½. We do this by setting equation (7) equal to one and solving for c_{fa}. This results in

$$\mathrm{MAC}_{WI} = 1 - \frac{(1 - c_{0.5a})}{2\sqrt{f(1-f)}}. \qquad (8)$$

Evaluations of equation (8) at representative values of f and $c_{0.5a}$ are given in table 3.

Finally, I present a formula for the variance of the treatment effect estimator when treatment effects differ for different clusters. Under this model expected value calculations are unchanged, but the variance of the estimated treatment difference $y_A - y_B$ under pseudo cluster randomization becomes

$$\mathrm{var}_{het} = \frac{2\sigma^2}{mn} \frac{f + w^2(1-f) + nq\{f - w(1-f)\}^2 + nq\omega\{f^2 + w^2(1-f)^2\}}{\{f + w(1-f)\}^2}. \qquad (9)$$

I note that under the heterogeneous treatment effects model there no longer exists a weight w that will result in a variance not depending on the unknown ρ and ω parameters. In the interest of space, no further calculations are presented with equation (9).

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

The current study shows how the results given in the paper by BMTP can be generalized to the case where there are heterogeneous treatment effects. It further shows that when comparing cluster randomized and randomized block designs in the presence of contamination, one must pay careful attention to whether statistical power is the criterion for evaluating the design or if mean squared error is the criterion. Finally, it is shown that in pseudo-cluster randomized

designs it is possible to create an estimator of the treatment effect whose variance does not depend on the unknown value of the ICC. One such design is the usual randomized block design. It is shown that one would choose to allocate an unequal number of individuals to treatment and control within each cluster only if the average contamination is expected to decrease substantially as a result. The further away from 1:1 allocation one goes, the more contamination must be reduced in order to prefer the unequal allocation.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Borm, G.F., Melis, R.J.F., Teerenstra, S. and Peer, P.G. (2005). Pseudo cluster randomization: a treatment allocation method to minimize contamination and selection bias. *Statistics in Medicine*, *24*, 3535-3547.

Cornfield, J. (1978). Randomization by Group: a formal analysis. *American Journal of Epidemiology*, *108*, 100-2.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. New York: Oxford University Press.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomization trials. *Psychological Methods, 2*, 173-85.

Rhoads, C. (forthcoming). The Implications of Contamination for Experimental Design in Education Research. *Journal of Educational and Behavioral Statistics*.

Schmidt, F.L. and Hunter, J.E. (1997). Eight Common but false objections to the discontinuation of significance testing in the analysis of research data. In L.A. Harlow, S.A. Mulaik and J.H. Steiger (Eds.) *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Ziliak, S.T. and McCloskey, D.N. (2004). Size Matters: the standard error of regressions in the American Economic Review. *The Journal of Socio-Economics*, 33, 527-546.

Appendix B. Tables and Figures

Not included in page count.

Table 1: $\mathrm{MAC}_{MSE,\,hom}$ values. Homogeneous treatment effects and MSE as evaluation criterion.

	ρ	0.001	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
$\mathrm{MAC}_{p,\mathrm{hom}}$	n=5	0.002	0.024	0.047	0.110	0.198	0.271	0.333	0.388	0.436	0.480	0.520
	n=25	0.012	0.106	0.186	0.343	0.486	0.570	0.629	0.673	0.708	0.737	0.762
	n=100	0.047	0.295	0.427	0.600	0.713	0.768	0.804	0.829	0.849	0.865	0.878
δ_T	m											
0.2	2	0.16	0.50	0.71	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	5	0.10	0.32	0.45	0.71	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	10	0.07	0.22	0.32	0.50	0.71	0.87	1.00	1.00	1.00	1.00	1.00
0.2	20	0.05	0.16	0.22	0.35	0.50	0.61	0.71	0.79	0.87	0.94	1.00
0.2	30	0.04	0.13	0.18	0.29	0.41	0.50	0.58	0.65	0.71	0.76	0.82
0.2	40	0.04	0.11	0.16	0.25	0.35	0.43	0.50	0.56	0.61	0.66	0.71
0.2	50	0.03	0.10	0.14	0.22	0.32	0.39	0.45	0.50	0.55	0.59	0.63
0.5	2	0.06	0.20	0.28	0.45	0.63	0.77	0.89	1.00	1.00	1.00	1.00
0.5	5	0.04	0.13	0.18	0.28	0.40	0.49	0.57	0.63	0.69	0.75	0.80
0.5	10	0.03	0.09	0.13	0.20	0.28	0.35	0.40	0.45	0.49	0.53	0.57
0.5	20	0.02	0.06	0.09	0.14	0.20	0.24	0.28	0.32	0.35	0.37	0.40
0.5	30	0.02	0.05	0.07	0.12	0.16	0.20	0.23	0.26	0.28	0.31	0.33
0.5	40	0.01	0.04	0.06	0.10	0.14	0.17	0.20	0.22	0.24	0.26	0.28
0.5	50	0.01	0.04	0.06	0.09	0.13	0.15	0.18	0.20	0.22	0.24	0.25
0.8	2	0.04	0.13	0.18	0.28	0.40	0.48	0.56	0.63	0.68	0.74	0.79
0.8	5	0.03	0.08	0.11	0.18	0.25	0.31	0.35	0.40	0.43	0.47	0.50
0.8	10	0.02	0.06	0.08	0.13	0.18	0.22	0.25	0.28	0.31	0.33	0.35
0.8	20	0.01	0.04	0.06	0.09	0.13	0.15	0.18	0.20	0.22	0.23	0.25
0.8	30	0.01	0.03	0.05	0.07	0.10	0.13	0.14	0.16	0.18	0.19	0.20
0.8	40	0.01	0.03	0.04	0.06	0.09	0.11	0.13	0.14	0.15	0.17	0.18
0.8	50	0.01	0.03	0.04	0.06	0.08	0.10	0.11	0.13	0.14	0.15	0.16

Table 2: $MAC_{MSE,het}$ values. Heterogeneous treatment effects and MSE as evaluation criterion.

	ρ	0.001			0.05			0.1			0.2			0.4		
	ω	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1
$MAC_{p,het}$	*n*=5	0.002	0.002	0.001	0.096	0.081	0.054	0.171	0.144	0.094	0.283	0.236	0.150	0.428	0.350	0.216
	n=25	0.011	0.009	0.006	0.291	0.242	0.154	0.403	0.330	0.205	0.504	0.405	0.246	0.582	0.459	0.273
	n=100	0.041	0.035	0.023	0.486	0.392	0.239	0.556	0.441	0.264	0.602	0.472	0.279	0.629	0.489	0.288
δ_T	*m*															
0.2	2	0.15	0.14	0.11	1.00	0.97	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	5	0.09	0.09	0.07	0.66	0.61	0.50	0.94	0.87	0.71	1.00	1.00	1.00	1.00	1.00	1.00
0.2	10	0.07	0.06	0.05	0.47	0.43	0.35	0.66	0.61	0.50	0.94	0.87	0.71	1.00	1.00	1.00
0.2	20	0.05	0.04	0.04	0.33	0.31	0.25	0.47	0.43	0.35	0.66	0.61	0.50	0.94	0.87	0.71
0.2	30	0.04	0.04	0.03	0.27	0.25	0.20	0.38	0.35	0.29	0.54	0.50	0.41	0.76	0.71	0.58
0.2	40	0.03	0.03	0.03	0.23	0.22	0.18	0.33	0.31	0.25	0.47	0.43	0.35	0.66	0.61	0.50
0.2	50	0.03	0.03	0.02	0.21	0.19	0.16	0.30	0.27	0.22	0.42	0.39	0.32	0.59	0.55	0.45
0.5	2	0.06	0.05	0.04	0.42	0.39	0.32	0.59	0.55	0.45	0.84	0.77	0.63	1.00	1.00	0.89
0.5	5	0.04	0.03	0.03	0.26	0.24	0.20	0.37	0.35	0.28	0.53	0.49	0.40	0.75	0.69	0.57
0.5	10	0.03	0.02	0.02	0.19	0.17	0.14	0.26	0.24	0.20	0.37	0.35	0.28	0.53	0.49	0.40
0.5	20	0.02	0.02	0.01	0.13	0.12	0.10	0.19	0.17	0.14	0.26	0.24	0.20	0.37	0.35	0.28
0.5	30	0.02	0.01	0.01	0.11	0.10	0.08	0.15	0.14	0.12	0.22	0.20	0.16	0.31	0.28	0.23
0.5	40	0.01	0.01	0.01	0.09	0.09	0.07	0.13	0.12	0.10	0.19	0.17	0.14	0.26	0.24	0.20
0.5	50	0.01	0.01	0.01	0.08	0.08	0.06	0.12	0.11	0.09	0.17	0.15	0.13	0.24	0.22	0.18
0.8	2	0.04	0.03	0.03	0.26	0.24	0.20	0.37	0.34	0.28	0.52	0.48	0.40	0.74	0.68	0.56
0.8	5	0.02	0.02	0.02	0.17	0.15	0.13	0.23	0.22	0.18	0.33	0.31	0.25	0.47	0.43	0.35
0.8	10	0.02	0.02	0.01	0.12	0.11	0.09	0.17	0.15	0.13	0.23	0.22	0.18	0.33	0.31	0.25
0.8	20	0.01	0.01	0.01	0.08	0.08	0.06	0.12	0.11	0.09	0.17	0.15	0.13	0.23	0.22	0.18
0.8	30	0.01	0.01	0.01	0.07	0.06	0.05	0.10	0.09	0.07	0.14	0.13	0.10	0.19	0.18	0.14
0.8	40	0.01	0.01	0.01	0.06	0.05	0.04	0.08	0.08	0.06	0.12	0.11	0.09	0.17	0.15	0.13
0.8	50	0.01	0.01	0.01	0.05	0.05	0.04	0.07	0.07	0.06	0.10	0.10	0.08	0.15	0.14	0.11

Table 3: Values of MAC_{WI} .

$c_{0.5a}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7
f=0.5	0.1	0.2	0.3	0.4	0.5	0.6	0.7
f=0.6	0.081	0.184	0.286	0.388	0.49	0.592	0.694
f=0.7	0.018	0.127	0.236	0.345	0.454	0.564	0.673
f=0.8	< 0	0	0.125	0.25	0.375	0.5	0.625
f=0.9	< 0	< 0	< 0	0	0.167	0.333	0.5
f=0.95	< 0	< 0	< 0	< 0	< 0	0.082	0.312